Proceedings of the 2002 International Conference on Auditory Display, Kyoto, Japan, July 2-5, 2002

DOLPHIN: THE DESIGN AND INITIAL EVALUATION OF MULTIMODAL FOCUS AND CONTEXT

David K McGookin

Department of Computing Science
University of Glasgow
Glasgow
Scotland
G12 8QQ
mcgookdk@dcs.gla.ac.uk
www.dcs.gla.ac.uk/~mcgookdk

Stephen A Brewster

Department of Computing Science
University of Glasgow
Glasgow
Scotland
G12 8QQ
stephen@dcs.gla.ac.uk
www.dcs.gla.ac.uk/~stephen

ABSTRACT

In this paper we describe a new focus and context visualisation technique called multimodal focus and context. This technique uses a hybrid visual and spatialised audio display space to overcome the limited visual displays of mobile devices. We demonstrate this technique by applying it to maps of theme parks. We present the results of an experiment comparing multimodal focus and context to a purely visual display technique. The results showed that neither system was significantly better than the other. We believe that this is due to issues involving the perception of multiple structured audio sources.

1. INTRODUCTION

Each year manufacturers are producing smaller and more powerful mobile computing devices. Palms, Pocket PC's and mobile phones have become ubiquitous. For example, 5.5 million mobile phones were sold in the UK in the 3 months before Christmas 2000 [1]. Manufacturers are now looking to produce multi-purpose mobile devices that will act as digital music players, mobile phones and web browsers.

Mobile computing is, however, very different from desktop computing. For example, the amount of screen resource available is only a fraction of that available on desktop computers. Also of great importance is the ability of users to employ their visual sense for safe navigation of the environment. For example, if you are checking your email on the move, you must split your visual attention between the reading of your mail and not falling down flights of stairs, getting run over by a car or any of the other dangers we can fall victim to by not looking where we are going. Even if we attempt to reduce these dangers by staying stationary, people could still walk into us, or a car could mount the pavement and hit us. In short, we need our eyes for much more important tasks than using a mobile computing device.

In an attempt to reduce the visual load on users we have designed a hybrid visual and spatialised audio focus and context visualisation technique called multimodal focus and context. Multimodal focus and context should not only increase the mobile device's display space, allowing more information to be displayed, but also reduce the demands on the user's visual sense by providing a constant audio context, allowing users to more quickly relocate where they are if and when their eyes are averted from the personal digital assistant (PDA) display. This should allow users to better and more safely navigate the physical environment.

In the remainder of this paper we will explain the relevant history of focus and context visualisation before describing multimodal focus and context. We shall then describe how data is represented in the spatialised audio space, before discussing the results of an experiment comparing multimodal focus and context to a purely visual technique.

2. FOCUS AND CONTEXT

Focus and context visualisation was originally, independently proposed by both Furnas [2] and Spence & Apperley [3]. Each of their proposed technique share the same common features but differ in key aspects.

All focus and context representations of information spaces share the same basic premise that more information is required to be presented than can be adequately, simultaneously, presented. To maximise the visual display space the information to be presented is split into two parts:

- **Focus:** That part of the information space that is of most interest to the user. This part is presented in maximum detail.
- **Context**: The rest of the information space. In order to allow all of the required information to be displayed this information is presented in much less detail than the focus.

The way in which the visual display is split between the focus and context largely determines whether the representation would be considered as Furnas's Fisheye [2] or Spence and Apperley's Bifocal Lens representation [3]. The Bifocal Lens has a much stricter visual disparity between the focus and context. In this system the focus and context can have different visual representations. For example, Spence and Apperley [3] demonstrated a visual bookshelf representation. Books were dragged from the bookshelf to another part of the screen where they were "opened" so that they could be read. Hence it is easy to tell if data is in the focus or the context. As was noted by Björk *et al.* [4], the Bifocal Lens style of focus and context means the data in the focus and context do not need to be the same.

There has been little research on applying focus and context to mobile computing devices. Notably, the work of Björk *et al.* [4], has attempted to apply Flip Zooming [5] focus and context visualisation to PDA's. In Flip Zooming the information space is broken into pages. Thumbnail representations of these pages are laid out, in order, on a grid. If a user wishes a better view of

one page, and hence make it the focus, they click the thumbnail. This causes the clicked page to expand whilst still retaining its relative position to the other pages. Björk applied this work to a personal contact manager [6] and a Web browser for PDA's [7]. However, this work still suffers from the issues previously outlined involving the demands on the visual sense.

3. MULTIMODAL FOCUS AND CONTEXT

Our new focus and context system augments the visual display with a new modality, specifically spatialized (3D) audio, to increase the available display area for information presentation. Because we use the visual display to represent the focus, whilst the audio space represents the context, we actually only use a transverse 2D audio plane (see Figure 1).

3.1. Overview

We decided to apply the Bifocal Lens concept to the multimodal display platform. There are several advantages to this approach. Firstly, as with the disparity between the focus and context on the bifocal display, there is a disparity between the visual and audio modalities. In other words, it is not possible to display visual representations visually in audio and *vice versa*. Another advantage is that the focus is high detail whereas the context is of lower detail. This fits well with the display platform in that it is not possible to display audio information in as much detail as visual information. These advantages mean that it is convenient to make the visual display the focus and the audio display the context. The splitting of the focus and context in this way should mean that the visual demand on the user is lowered and that they will be able to retain position in the map even when their visual attention is distracted by environmental stimuli.

3.2. Fitting together the focus and context

The focus essentially "floats" over the context. Users see the focus on the PDA screen. The data which are to the right and forward from the focus are 'played' in the audio space, to the right and forward of the user. The data that are to the left and rear of the focus, are played to the left and rear of the user (see Figure 1).

Users navigate through the space via scrollbars on the visual display. The act of moving a part of the display from the focus to the context actually means moving map items from the visual to the audio modality. When this occurs the visual representation of the map item is replaced with a spatialised audio representation. For example, scrolling to the right will cause the left part of the focus to move from the visual display to the audio display (and hence move from the focus to the context).

Audio representations of map items remain the same relative distance from each other as when they are displayed in the visual modality. In essence, we are moving a lens (the visual display) over a large information space. The data that the visual display is over are represented visually; the rest of the information space is represented in audio.

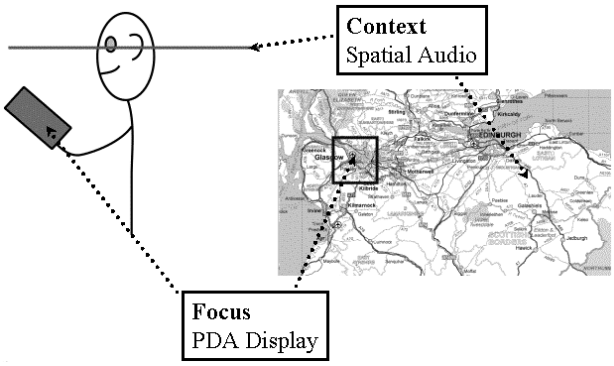

Figure 1. *Overview of multimodal focus and context.*

4. DESIGNING THE CONTEXT

To properly explain the rest of multimodal focus and context we shall use the presentation of theme park visitor maps on PDA's as an example.

By their very nature theme parks are large and thus difficult to navigate. Most of the visitors will never have visited the park before, they have a limited time at the park, and entry to the park will have cost a lot of money. It is therefore important for the visitor to be able to quickly and effectively navigate the park. Hence visitors use maps. However, visitors must also be aware of what is around them due to the dangers of the real world environment previously outlined. These features make theme park maps a good candidate to apply multimodal focus and context to.

We shall describe the audio part of our design in several stages. We will start by describing the individual audio cues that we use to represent rides before describing how the audio space is managed.

4.1. Audio Cues

To display the theme park rides in the audio space, we first must decide the attributes to be communicated. We decided that a typical user might wish to know the type of ride (e.g. a roller coaster, water ride, etc.), how intense the ride was and how much the ride would cost. These attributes as well as their values are given in Table 1 below:

Attribute	Description
Type	This attribute categorises the ride into one of three types. Rollercoaster, Water Ride or Static Ride.
Intensity	The intensity is either one of low, medium or high. Large, fast, rollercoasters would be an example of high intensity rides.
Cost	Cost can either be one of low, medium or high.

Table 1. *Attributes encoded into the audio cues.*

These attributes were represented in audio by encoding them into Earcons [8]. Earcons are short structured audio messages, which can be effectively used to convey such information [9].

In order to represent the above attributes we have used a variant of the hierarchical Earcon type [8]. Here we map each of the attributes to a separate auditory parameter. The mapping of parameters was done in line with the observations of Norman [10] on visual mappings, and the Earcons were designed in accordance with the guidelines of Brewster *et al.* [11]. The Earcon structure is described in Table 2. The Earcons were constructed using the Cakewalk MIDI sequencer and were recorded as .wav files from a Roland Super JV-1080 synthesiser for use in the spatialisation system.

4.2. Placing the Sounds

There are several cues that the human auditory system uses to localise audio sources. These cues can be encoded into a head related transfer function (HRTF). An HRTF is in essence a function, which takes an audio source and a position, and filters the audio source such that it is perceived to come from the supplied location [13].

Auditory Parameter	Use
Timbre	As ride type is a substitutive scale [10], i.e. we cannot say that a roller coaster is greater than a water ride, we have mapped this to timbre. We have taken care to ensure that we choose obviously different instruments. We use a trumpet to represent a rollercoaster, a banjo to represent a water ride and a piano to represent a static ride.
Rhythm	As this is an additive scale, we have mapped it to the intensity attribute. Three distinct rhythms were used representing low, medium and high intensity. In accordance with the guidelines of Brewster *et al.* [11], we used a varying number of notes to help differentiate the rhythms, with 2, 4 and 6 notes used respectively for low, medium and high intensities
Pitch	We mapped the cost of a ride to pitch, with a higher pitch representing a greater cost. As absolute pitch perception is difficult for most people, we ensured that there was a gross difference (at least an octave) between the pitches. In addition we altered the absolute position of each Earcon within an octave to provide more variation [11].

Table 2. *Mapping of ride attributes to auditory parameters.*

Most current personal computer (PC) sound cards are supplied with generalised HRTFs which are accessible via the Microsoft DirectX API. We have used the HRTF on the Videologic Sonic Fury sound card (this card is marketed as the Turtle Beach SantaCruz in the USA) which also combines features from Sensaura to provide a more realistic near field effect. The audio was presented through Sennheiser HD-25 headphones.

4.3. Audio Overload

One of the problems with the system so far outlined is that there will be a much greater amount of audio information to be presented than visual information. For example in the experimental version we will shortly describe, there were 27 individual rides and only 3-4 of them could be represented on the visual display. Twenty three audio sources simultaneously playing is clearly much more than a user can handle, and it became clear, during formative testing, that some way to reduce the audio whilst still retaining the ability to use it to navigate the theme park map was important. We developed a system called 'priority zones' to provide a framework for the rule-based reduction of the amount of audio. Priority zones borrow many of the ideas of the Degree of Interest (DOI) function of Furnas's original fisheye concept [2]. The idea is that less important things that are far away should be given less display resource than closer, more important things. Far away, but very important things should have more resource than very unimportant but close things. It is simple, in the visual domain, to determine what is meant by using "less resource" to display information, we simply reduce the size of the visual icon. In the audio domain determining what "less resource" means is more difficult. We considered using the technique employed in Sawney and Schmandt's Nomadic Radio [12] personal notification system. Here, more important messages were played using more detailed audio means. For example, for low priority messages auditory icons were used, whereas for high importance messages, speech was used. We decided against this approach because we believe there will be many more sounds playing concurrently in our system than in Nomadic Radio. Because of the amount of audio, we were interested in looking at the more extreme solution to the problem of audio overload which is to completely switch off audio that is not required. Using the Earcon representation, it does not make sense to reduce the number of parameters represented by removing the pitch, timbre or rhythm of a sound. We also considered reducing the volume at which an Earcon was presented. This is a direct analogy with the reduction of a visual stimuli size. However the volume of a sound is an important cue to its distance, particularly when the object does not come from a natural source [13]. Reducing the volume is likely to confuse the user over the distances of objects.

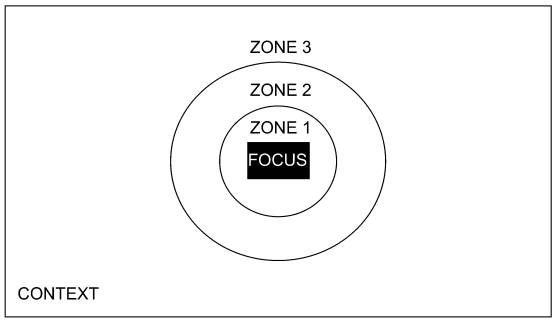

Figure 2. *Relationship of priority zones to the focus and context.*

In our system we give each of the rides (represented by an Earcon in the audio space) a priority number between 1 and 3 which specifies its "importance". The lower the number the less important the ride. Numbers were allocated based on the highest attribute between the cost and intensity attributes. Therefore a low cost, low intensity ride would be allocated a

priority number of 1, whereas a low cost, high intensity ride would be allocated a priority of 3.

Extending out from the focus, and fixed relative to it, in concentric circles, are the priority zones (see Figure 2). For a sound (representing a ride) to be played, it must lie in a priority zone with a number less than or equal to its own number. This means that sounds are switched on and off dynamically as they move between zones. In doing this we can remove those audio sources that are unlikely to be important based on the user's current map location. For example Figure 3 represents the 2D planar audio space for a particular map. The focus (which is represented visually on a PDA screen) is at the center.

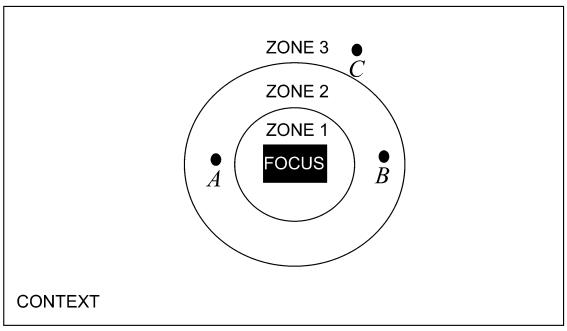

Figure 3. *Example of the audio space for a given map with three Earcons.*

This particular map contains three Earcons, A, B and C. Earcon A represents a low intensity, low cost ride, Earcon B a medium intensity, low cost ride, and Earcon C a low intensity, high cost, ride. According to our previously outlined system for allocating priority numbers, Earcon A will have a priority number of 1, Earcon B will have a priority number of 2 and Earcon C will have the priority number of 3. Therefore in this map Earcons B and C will be audible to the user since they are lying in a priority zone with a number less than or equal to their own. Earcon A lies in priority zone 2 and since it has the priority number 1, it will not be "played". Figure 4 shows the same map after the user has moved the focus position by "scrolling" the visual display. As the priority zones are fixed relative to the focus they also move. Here, Earcon A will be played as it has moved from priority zone 2 to priority zone 1. However Earcon B has moved from priority zone 2 to priority zone 3 and will stop playing. Earcon C has not switched zones so will continue to be played.

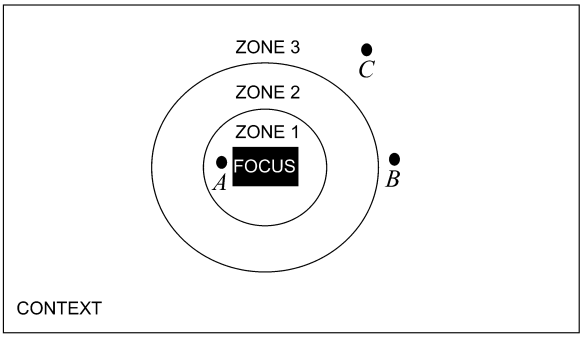

Figure 4. *Example of the location of priority zones after the user has moved the focus.*

One of the problems with priority zones is setting their boundaries. That is, when should classes of sounds be switched on and off? We have found that this is a non-trivial problem as users must have enough information to aid navigation, but not so much that the audio overloads the user whilst navigating. In our experiment we have attempted to push more towards a reduction in annoyance, as we do not know how much information is required in audio to enable effective navigation.

5. EVALUATION AND RESULTS

To determine the effectiveness of the multimodal focus and context system, called Dolphin, outlined above, we evaluated it against a standard scrolling view. The standard scrolling view is the same as multimodal focus and context except there is no audio. Whilst it would have been preferable to evaluate against a purely visual focus and context technique, there has been little work to show the effectiveness of visual focus and context. Also, scrolling views are the most popular way to present large information spaces on smaller screens.

Sixteen people participated in the experiment, all of whom were students at the Computing Science Department of Glasgow University, and therefore, were experienced computer users. There were two conditions; the multimodal focus and context condition and the visual scrolling display condition. The experiment was of a within groups design. The order of the conditions was counterbalanced to avoid learning effects. Due to the limitations of audio on current mobile computing devices, the experiment was run in a 6x6 cm window on a standard desktop machine.

Before performing the experiment participants were first given training. The training consisted of two parts. In the first part, participants were trained on the icons they would be exposed to in the experiment. Participants were given a sheet describing how the icons were constructed, before being allowed 5 minutes to familiarise themselves with a Web page, containing all of the appropriate icons used in the experiment. Participants were then presented with three of the icons independently and asked to describe what they were. If a participant failed to correctly identify more than one attribute on any test icon he/she would be given another 5 minutes to refamilarise themself with the web page before retesting. Earcon training was similar to that for the icons. Once the participant had successfully completed the first part of the training, he/she was given a sheet which explained all of the features of the experimental set-up, before attempting a shortened version of the appropriate experimental condition. This provided an opportunity for participants to ask questions as well as familiarise themselves with the task to be performed.

In the experiment, participants were asked to create routes around fictional, standardised, theme park maps. E.g. "Create a minimum route around all of the high intensity water rides". Note that in all cases the participant was asked about 2 attributes, the type of ride as well as either the intensity or cost of the ride. Participants were also never told how many rides of a particular type there were in the map, as we wanted to use the fact that they missed rides as an indication of how well they had understood the map in that condition. The icons used to represent theme park rides in the visual condition were based around a similar abstract technique as the Earcons described earlier. Type was specified as shape, cost as the number of dots on the shape and intensity as the shade of the dots on the shape. It would have been possible to use pictorial images to represent rides visually, however it would be difficult to represent parameters such as cost or intensity in a pictorial representation of a ride.

Figure 5 shows a screenshot of the visual scrolling condition (which also represents the visual display of the multimodal focus and context condition) showing a medium intensity, medium cost static ride (the square), and a high intensity, low cost water ride (the circle). When a participant found a ride that should be added he/she clicked the small black square in the centre of the icon to add it to his/her tour.

5.1. Hypotheses

There were three main hypotheses investigated. That participants would take less time to complete a tour in the multimodal focus and context condition, participants would make overall shorter routes in the multimodal focus and context condition, and that there would be less occasions in the multimodal focus and context condition where participants would miss out one or more of the rides that should have been added to the route, or added rides which should not have been added to the route. The main purpose of these hypotheses was to try to measure how well participants understood the overall map.

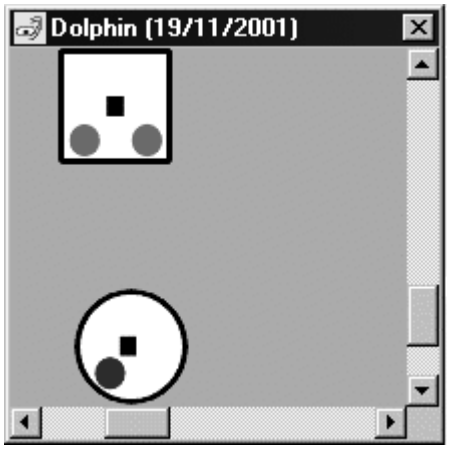

Figure 5. *A screenshot of the visual interface to dolphin.*

5.2. Results

Two-tailed T tests were performed on the results for the three hypothesis mentioned above. Whilst none of the results of these tests showed a significant difference between the two conditions, we believe that it is likely that the spatial audio used in the multimodal focus and context system both assisted and confused the user in equal measure. That is, in some situations the participant successfully used the audio to identify where he/she was, or where the next ride to be added to his/her route was. On other occasions, however, the audio was annoying or caused the participant to misinterpret his/her next direction. We can at this stage only speculate as to the actual causes for the problems with the audio space. Whilst we have followed the guidelines for the construction of the Earcons [11], these guidelines have been based on non-spatialised presentations of single Earcons. They do not refer to spatialised placement, or multiple concurrent occurrences of Earcons. Almost all of the research into the limits of spatialisation, the minimum audible angle (MAA) [14, 15], stream analysis [16] and so forth deals with either noise, speech or long musical compositions. We have identified therefore, that there is a lack of research into the limits of extracting information from multiple, spatialised, structured audio sources. For example, we have no evidence to show the number of Earcons that can be simultaneously presented, or how far apart these Earcons must be, for the information contained within them to be reliably extracted. Therefore, we intend to investigate the issues surrounding the spatialisation of multiple structured audio cues and feed the results back into our multimodal focus and context system.

6. CONCLUSIONS

We have presented a technique for increasing the display space of mobile devices by augmenting the visual display with a spatial audio representation. This technique uses the principles of focus and context information visualisation to link together both of these displays. How information is represented in both the visual and audio displays has been explained.

Multimodal focus and context has been evaluated against a purely visual scrolling view with standardised theme park maps. There was no significant difference in either the accuracy or speed of navigation between the two conditions. We believe this is due, in part, to the lack of information for the creation of spatialised audio spaces which are populated with structured audio. Future research into these issues will be applied back to Dolphin to determine the performance gain they provide. We believe that with some further development, multimodal focus and context provides a strong candidate to increase the display space, and lower the visual load on users of PDAs.

7. ACKNOWLEDGEMENTS

This work was supported by EPSRC studentship 00305222.

8. REFERENCES

[1] BBC News, "http://news.bbc.co.uk/hi/english/business/newsid1100000/1100250.stm," 2000.

[2] G. W. Furnas, "Generalized Fisheye Views," presented at CHI'86, Boston, MA, 1986, pp. 16-23.

[3] R. Spence and M. D. Apperley, "Database navigation: An office environment for the professional," *Behaviour and Information Technology*, vol. 1, pp. 43-54, 1982.

[4] S. Björk and J. Redström, "Redefining the Focus and Context of Focus+Context Visualizations," presented at IEEE Symposium on Information Visualization 2000, 2000.

[5] L. E. Holmquist, "Focus+Context Visualization with Flip Zooming and the Zoom Browser," presented at CHI'97, Atlanta, Georgia, 1997, pp. 263-264.

[6] S. Björk, J. Redström, P. Ljungstrand, and L. E. Holmquist, "PowerView: Using Information Links and Information Views to Navigate and Visualize Information on Small Displays," presented at Handheld and Ubiquitous Computing 2000, Bristol, UK, 2000, pp. 45-62.

[7] S. Björk, L. E. Holmquist, J. Redström, I. Bretan, R. Danielsson, J. Karlgren, and K. Franzen, "WEST: A Web Browser for Small Terminals," presented at UIST'99, Asheville, NC, 1999, pp.187-196.

[8] M. M. Blattner, D. A. Sumikawa, and R. M. Greenberg, "Earcons and Icons: Their Structure and

Common Design Principles," *Human Computer Interaction*, vol. 4, pp. 11-44, 1989.

[9] S. A. Brewster, "Providing a structured method for integrating non-speech audio into human-computer interfaces," in *Department of Computer Science*. York: University of York, 1994, pp. 265.

[10] D. A. Norman, "Cognitive Artifacts," in *Designing Interaction: Psychology at the Human-Computer Interface*, vol. 1, *Cambridge Series on Human-Computer Interaction*, J. M. Carroll, Ed. Cambridge: Cambridge University Press, 1991, pp. 17-38.

[11] S. A. Brewster, P. C. Wright, and A. D. N. Edwards, "Experimentally derived guidelines for the creation of earcons," presented at HCI'95, Huddersfield, 1995, pp.155-159.

[12] N. Sawhney and C. Schmandt, "Nomadic Radio: Speech & Audio Interaction for Contextual Messaging in Nomadic Environments," *ACM Transactions on CHI*, vol. 7, pp. 353-383, 2000.

[13] W. W. Gaver, "Auditory Interfaces," in *Handbook of Human-Computer Interaction*, vol. 1, M. G. Helander, T. K. Landauer, and P. V. Prabhu, Eds., 2nd ed. Amsterdam: Elsevier, 1997, pp. 1003-1041.

[14] S. A. Gelfand, *Hearing: An introduction to psychological and physiological acoustics*. New York: Marcel Dekker, 1981.

[15] B. C. J. Moore, *An Introduction to the psychology of hearing*, 4th ed. London: Academic Press, 1997.

[16] A. S. Bregman, *Auditory Scene Analysis*, vol. 1, 1 ed. London, England: MIT, 1994.